Linear Regression (Hồi quy tuyến tính)

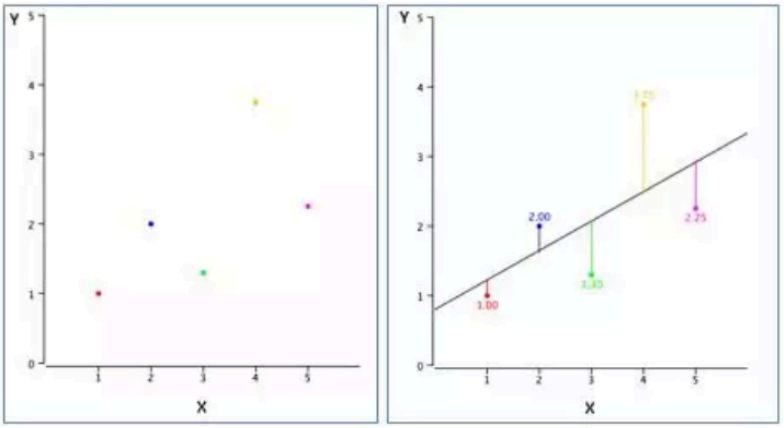
Hoàng Vũ Minh

Linear Regression hay hồi quy tuyến tính là một kỹ thuật học có giám sát, phân tích quan hệ giữa biến phụ thuộc Y với 1 hay nhiều biến độc lập x, dựa trên các dữ liệu có gán nhãn cho trước.

1. *Ý tưởng cơ bản:*

Hiểu một cách đơn giản bằng hình học: hồi quy tuyến tính tương đương với việc ta tìm một đường thẳng sao cho khoảng cách từ nó đến các điểm dữ liệu là nhỏ nhất (tức là nó sát với các điểm dữ liệu nhất), như vậy ta có thể dùng đường đó để dự đoán các điểm xuất hiện trong tương lai

Ví dụ trong trường hợp hồi quy tuyến tính đơn biến (chỉ có 1 biến):



Trong trường hợp đa biến, thay vì đường thẳng, nó sẽ là các mặt phẳng. Ta có thể dự đoán hệ số tương quan (rô) của chúng xấp xỉ 1 nếu như sau khi plot ra chúng gần như tạo thành 1 đường thẳng, tức là phụ thuộc tuyến tính cực mạnh.

2. *Phương trình cho trường hợp đơn biến:*

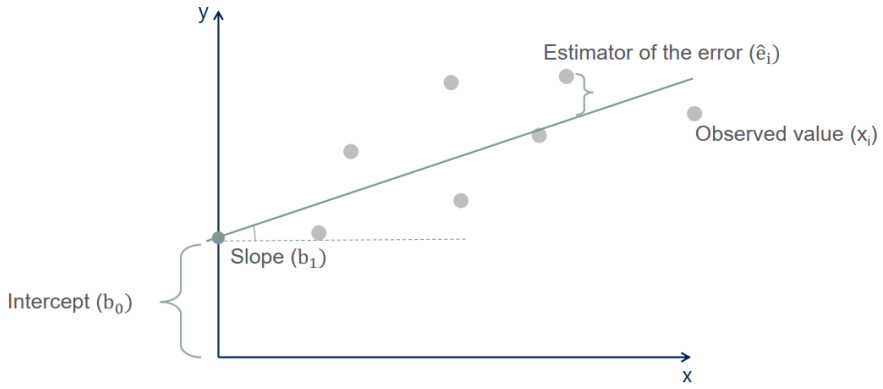
$$Y_i = \beta 0 + \beta 1 * X_i$$

Trong đó:

- Y_i là biến phụ thuộc
- β0 là constant/intercept, hay hệ số tự do

- β1 là slope (độ đốc), mô tả sự biến đổi của biến phụ thuộc Y.
- X_i là các biến độc lập

Dưới đây là mô tả hình ảnh về linear regression (nguồn 365 data science).



3. *OLS (ordinary least squares)*

Như vậy, ta cần tìm B0 và B1 sao cho giá trị dự đoán và giá trị thực tế sai khác ở mức tối thiểu. Phương pháp: OLS - bình phương nhỏ nhất có thể được sử dụng để tính toán ra B0 và B1 trong trường hợp này. Nghiệm khi sử dụng:

$$\beta 1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\beta 0 = \bar{y} - B_1\bar{x}$$

Với $\bar{x}$ là các giá trị trung bình của x, được tính bằng: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Với $\bar{y}$ là các giá trị trung bình của x, được tính bằng: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

-

4. *Hàm mất mát:*

Ta có thể sử dụng MSE (mean square error) làm hàm mất mát cho hồi quy tuyến tính. MSE sẽ đo trung bình bình phương sự chênh lệch giữa giá trị thực tế và giá trị dự đoán:

$$MSE = \frac{1}{n}\sum(Y_i - \widehat{y_i})^2$$

Trong đó $\widehat{y}$ là các giá trị dự đoán, y là các giá trị thực tế. Như vậy có thể nhận xét, MSE càng nhỏ thì giá trị dự đoán của mô hình các sát với thực tế.

5. *Độ đo cho các bài toán hồi quy*

Cho các bài toán hồi quy, ta thường sử dụng metric RMSE và R2-square.

4.1. RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_i - \widehat{y_i})^2}$$

RMSE có ý nghĩa tương tự với MSE, càng nhỏ thì chứng tỏ giá trị dự đoán càng sát với giá trị thực tế.

4.2. R2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Trong đó:

- Yi là giá trị thực tế
- $\widehat{Y}$ là giá trị dự đoán
- $\overline{y}$ là giá trị trung bình của biến phụ thuộc

R2 đo lường tỉ lệ biến thiên của Y có thể dược giải thích bằng X, nó nằm trong khoảng 0 đến 1. Tức là nếu R2 càng cao, càng gần 1 thì chứng tỏ biến phụ thuộc Y càng được biểu diễn tốt bằng x và ngược lại.

6. *Nhược điểm của hồi quy tuyến tính:*

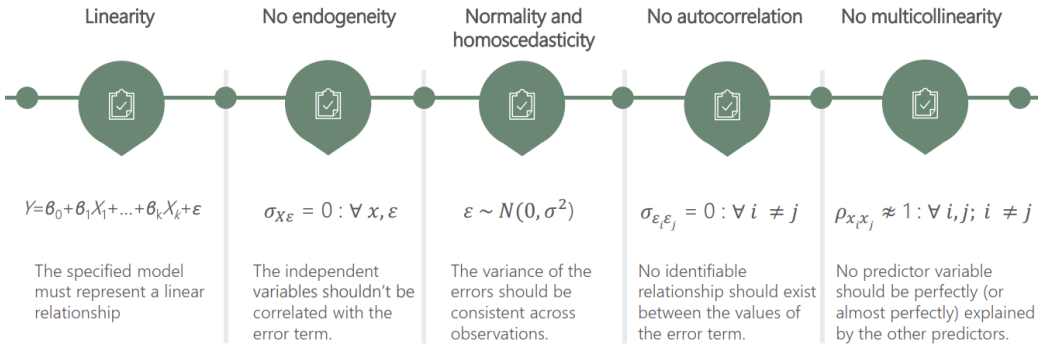
Mặc dù đơn giản và dễ sử dụng nhưng hồi quy tuyến tính tồn tại một số nhược điểm lớn:

- Nhạy cảm với các noise và outliers, những điểm này có thể làm cho kết quả của mô hình tệ đi rất nhiều, vì vậy để mô hình hoạt động tốt, cần xem xét loại bỏ các outliers
- Các biến độc lập có mối quan hệ tuyến tính mạnh với nhau: hiện tượng này có thể gọi là đa cộng tuyến, có thể gây ra ảnh hưởng rất lớn đến ước lượng hệ số hồi quy.
- Không diễn giải được các mối quan hệ phi tuyến

-> cách khắc phục các nhược điểm: tiền xử lý và kiểm tra trước khi xây dựng mô hình:

- Kiểm tra tính tuyến tính của dữ liệu, với trường hợp đơn biến, đơn giản là có thể vẽ đồ thị (plot) ra và xem nó có dạng đường thẳng hay không.
- Kiểm tra sự tương quan giữa các biến độc lập trước khi tiến hành hồi quy.

7. Mấy cái giả định của hồi quy tuyến tính và cách check



Dưới đây là một số giả định, việc đáp ứng các giả định này giúp ta đảm bảo kết quả hồi quy tìm được là chuẩn xác (đúng hơn đây là giả định để OLS hoạt động tốt).

Linearity (tính tuyến tính): Cần mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc, nếu không nó sẽ không work. Cách check:

- Vẽ scatter plot cho biến phụ thuộc với từng biến độc lập, nếu xu hướng các điểm tập trugn thành một đường thẳng thì có lẽ chúng có mqh tuyến tính.

No endogeinity (không có tính nội sinh): các biến độc lập không được có sự tương quan với error term. Hay nói cách khác, không có sự tương quan giữa các biến độc lập với sai số. Cov(x,e) = 0

- Nguyên nhân: một trong số lý do thường thấy là do thiếu biến liên quan (omitted variable bias).
    - Nó như kiểu vdu y = x1 + x2, mà ta biểu diễn nhầm chỉ thành y = x1. Vì y được giải thích bằng cả x1 và x2 nên phần nào đó 2 biến độc lập này cũng có một chút quan hệ (vì nó cùng lý giải y).
    - Những gì không được lý giải bởi model sẽ bị đưa vào sai số, cho nên như thế sai số mới có tương quan với biến độc lập x1 (vì trong sai số có cả phần của x2).

Normality and Homoscedasticity: Sai số tuân theo phân phối chuẩn (0, xm2) và phương sai của sai số

- Sai số theo phân phối chuẩn: có thể coi luôn là đúng khi cỡ mẫu lớn (đly giới hạn trung tâm)
- Kỳ vọng = 0: gần như luôn thỏa mãn do có intercept
- Sai số phải có phương sai không đổi giữa các quan sát (đồng nhất). Phản ví dụ: sự thay đổi chi tiêu của ng nghèo thường ít nhưng người giàu thường hên xui.
    - Để ngăn cái này: chú ý check OVB và xử lý outliers
    - Log transform: Semi-log model: Khi chỉ áp dụng log lên Y hoặc X. Log - log là cả 2.

No autocorrelation: Không có sự tương quan giữa các sai số, tức là Cov(ei, ej) = 0 với mọi i j.

- Cách check: dùng một số kiểm định hoặc vẽ ACF và phân tích.

No multicollinearity: Không có đa cộng tuyến

- Đa cộng tuyến được hiểu là khi 2 biến phụ thuộc có liên quan mạnh với nhau (kiểu chúng có thể được biểu diễn hoàn toàn bởi nhau).
- Cách check: vẽ heatmap, correlation matrix.